The disconcerting potential of online disinformation: persuasive effects of astroturfing comments and three strategies for inoculation against them

Zerback, Thomas ; Töpfl, Florian ; Knöpfle, Maria

Abstract: This study is the first to scrutinize the psychological effects of online astroturfing in the context of Russia's digitally enabled foreign propaganda. Online astroturfing is a communicative strategy that uses websites, "sock puppets," or social bots to create the false impression that a particular opinion has widespread public support. We exposed N = 2353 subjects to pro-Russian astroturfing comments and tested: (1) their effects on political opinions and opinion certainty and (2) the efficiency of three inoculation strategies to prevent these effects. All effects were investigated across three issues and from a short- and long-term perspective. Results show that astroturfing comments can indeed alter recipients' opinions, and increase uncertainty, even when subjects are inoculated before exposure. We found exclusively short-term effects of only one inoculation strategy (refutational-same). As these findings imply, preemptive media literacy campaigns should deploy (1) continuous rather than one-time efforts and (2) issue specific rather than abstract inoculation messages.

DOI: https://doi.org/10.1177/1461444820908530

The disconcerting potential of online disinformation: Persuasive effects of astroturfing comments and three strategies for inoculation against them

Journal:	*New Media and Society*
Manuscript ID	NMS-19-0659.R2
Manuscript Type:	Original Manuscript
Keywords:	disinformation, misinformation, Russia, state propaganda, online astroturfing, opinion certainty, uncertainty, countermeasures, inoculation
Abstract:	This study is the first to scrutinize the psychological effects of online astroturfing in the context of Russia's digitally-enabled foreign propaganda. Online astroturfing is a communicative strategy that uses websites, "sock puppets," or social bots to create the false impression that a particular opinion has widespread public support. We exposed N = 2,353 subjects to pro-Russian astroturfing comments and tested: (1) their effects on political opinions and opinion certainty, and (2) the efficiency of three inoculation strategies to prevent these effects. All effects were investigated across three issues and from a short- and long-term perspective. Results show that astroturfing comments can indeed alter recipients' opinions, and increase uncertainty, even when subjects are inoculated before exposure. We found exclusively short-term effects of only one inoculation strategy (refutational-same). As these findings imply, preemptive media literacy campaigns should deploy (1) continuous rather than one-time efforts and (2) issue-specific rather than abstract inoculation messages.

Abstract

This study is the first to scrutinize the psychological effects of online astroturfing in the context of Russia's digitally-enabled foreign propaganda. Online astroturfing is a communicative strategy that uses websites, "sock puppets," or social bots to create the false impression that a particular opinion has widespread public support. We exposed $N = 2{,}353$ subjects to pro-Russian astroturfing comments and tested: (1) their effects on political opinions and opinion certainty, and (2) the efficiency of three inoculation strategies to prevent these effects. All effects were investigated across three issues and from a short- and long-term perspective. Results show that astroturfing comments can indeed alter recipients' opinions, and increase uncertainty, even when subjects are inoculated before exposure. We found exclusively short-term effects of only one inoculation strategy (refutational-same). As these findings imply, preemptive media literacy campaigns should deploy (1) continuous rather than one-time efforts and (2) issue-specific rather than abstract inoculation messages.

Keywords: disinformation, misinformation, Russia, state propaganda, online astroturfing, opinion certainty, uncertainty, countermeasures, inoculation

The disconcerting potential of online disinformation: Persuasive effects of astroturfing comments and three strategies for inoculation against them

Particularly in the aftermath of the 2016 US national election, disinformation and its consequences for democratic societies have been subject to extensive political (European Commission, 2018) and scholarly debate (e.g., Bennett & Livingston, 2018). At the most abstract level, disinformation can be understood as "[i]naccurate or manipulated information / content that is spread intentionally. This can include false news, or involve more subtle methods such as false flag operations, feeding inaccurate quotes or stories to innocent intermediaries, or knowingly amplifying biased or misleading information" (Weedon, Nuland, & Stamos, 2017, p. 5). Therefore, disinformation is also persuasive communication (Zhang, Carpenter, & Ko, 2013). In this paper, we deal with an important and widespread subtype of disinformation, known as "astroturfing" (Kovic, Rauchfleisch, Sele, & Caspar, 2018; Zhang et al., 2013). Astroturfing can be defined as the "manipulative use of media and other political techniques to create the perception of a grassroots community organization where none exists for the purpose of political gain" (McNutt & Boland, 2007, p. 169). Although the phenomenon is not new, the Internet and especially social media have paved the way for new forms, often referred to as digital or online astroturfing (Kovic et al., 2018; Zhang et al., 2013).

A central strategic instrument of online astroturfing is the manufacturing of user comments designed to appear as authentic citizen voices on highly visible news or social networking sites (SNS). We focus here on this specific form of online astroturfing because it has been one of the most widely debated in the context of national elections across the Western world (Ferrara, 2017; Kovic et al., 2018; Zelenkauskaite & Balduccini, 2017). Examples of targeted campaigns include the 2016 presidential election in the US (Bessi & Ferrara, 2016; Woolley & Guilbeault, 2017), the 2017 presidential election in France (Ferrara, 2017), and the 2012 presidential elections in South Korea (Keller, Schoch, Stier, &

Yang, 2019). As a key sponsor of these astroturfing activities, various authors have pointed to Russia's ruling elites (see, for instance, Bugorkova, 2015; Zelenkauskaite & Balduccini, 2017), who are closely tied to an organization known as the Internet Research Agency (IRA) or Russia's "troll factory" (Lysenko & Brooks, 2018; Ruck, Rice, Borycz, & Bentley, 2019). In 2013, this entity employed approximately 600 people with an estimated annual budget of US$ 10 million (Bugorkova, 2015). Amongst others, it targeted foreign audiences by setting up fake SNS accounts mimicking grassroots support for Russian policies on a range of news and social-media platforms (Bennett & Livingston, 2018; Kovic et al., 2018).

Among Western political leaders, these digitally enabled propaganda efforts have sparked not only concern but explicit indignation (European Commission, 2018). In academia, they have stimulated a fast-growing body of research on the phenomenon. So far, however, this research has focused almost exclusively on how to identify fake accounts or automated social bots (Keller et al., 2019; King, Pan, & Roberts, 2017). By contrast, we still know very little about the psychological effects that such manufactured user comments exert on media audiences, and even less about possible ways of preventing them. Against this background, our study advances existing research in three ways:

(1) We examine whether online astroturfing comments affect the political opinions and opinion-certainty of those exposed to them.

(2) We investigate whether these persuasive effects can be mitigated, or even prevented, by inoculation messages designed to educate the audience about the manipulative intent and argumentative tactics of the astroturfing actors.

(3) We analyze the duration of the inoculation's immunizing effects.

Our study is based on a three-wave experiment (N = 2,353) carried out over the course of four weeks within the sociopolitical context of Germany. Participants were exposed to typical online astroturfing comments posted beneath social media news items and dealing with one of three issues prone to Russian astroturfing activities: the poisoning of former

Russian intelligence officer Sergei Skripal, the manipulation of the 2016 US presidential election, and the use of toxic gas by a close Russian ally, the Syrian government. All issues had been among the top stories on the Germany news agenda for several days. After exposure, we tested the comments' persuasive effects, as well as the short and long-term efficiency of three different inoculation treatments in countering them.

The effects of astroturfing comments on personal opinions

Online astroturfing comments imitate ordinary citizens' voices in order to create the impression that a certain opinion has widespread public support, while the real agent behind the message conceals his identity (Zhang et al., 2013). They are almost impossible to distinguish from authentic user comments; hence the audience find themselves in situations where they are either completely unaware of the fact that a comment might be sponsored by a principal, or they may suspect such an influence but cannot be entirely sure about it. Given their authentic appearance and the lack of knowledge, and/or uncertainty, on the part of audiences, astroturfing comments carry the potential to influence the opinions of those who read them.

An answer to the question, how astroturfing comments can alter personal opinions is given by exemplification research, which investigates the effects of ordinary citizen depictions in the media (also known as "exemplars") (Zillmann, 1999). Exemplars possess several characteristics contributing to their persuasive potential: firstly, as personalized information they attract the audience's attention, making persuasive effects more likely in the first place (Taylor & Thompson, 1982). Secondly, the opinion voiced by an exemplar becomes cognitively available and accessible in the recipients' memories (Zillmann, 1999), and thus has a greater chance of influencing subsequent judgments (Domke, Shah, & Wackman, 1998). Finally, fellow citizens are often considered to be more trustworthy and similar to ourselves by comparison with other actors in the media, such as e.g. politicians

(Lefevere, Swert, & Walgrave, 2012). Trustworthiness and similarity have both been shown to be strong facilitators of persuasive effects (Hovland, Janis, & Kelley, 1953).

Although, depictions of citizens seem to hold great potential to influence the opinions of those confronted with them, empirical evidence is rather mixed. Whereas some researchers have observed opinion changes resulting from exemplar exposure in traditional (e.g., Daschmann, 2000) and online media (e.g., Sikorski, 2018), others could not find such effects (e.g., Zerback & Peter, 2018). This leads to the question of why online astroturfing comments, in particular, should exert a persuasive influence. The answer lies in the way they are composed: in many cases, astroturfing comments do not merely consist of an opinion, but also include arguments that support the position advocated. An analysis by the EU vs. Disinformation project (2019) found that, particularly in the case of Russian propaganda, a common strategy was to offer alternative explanations for negative events for which Russia was being publicly accused. These pro-Russian astroturfing messages deny Russian responsibility, present alternative culprits, or portray Russia as the victim of unfounded Russophobia or public persecution (see also Nimmo, 2015). Persuasion research has repeatedly shown that arguments included in a message increase its persuasive impact (Petty & Cacioppo, 1984), which should also apply to astroturfing comments.

So far, only two studies have provided insights into the effects of astroturfing activities on audience attitudes, although in all these cases, researchers have not used online comments but other types of astroturfing information. In an experiment, Cho, Martens, Kim, and Rodrigue (2011) showed that people who were exposed to astroturf websites became more uncertain, as compared with those who saw real grassroots websites, about the causes of global warming and humans' role in the phenomenon. Interestingly, these effects occurred despite the fact that participants had (correctly) perceived the information from the astroturfing websites to be less credible and the organization less trustworthy. In another study, Pfau, Haigh, Sims, and Wigley (2007) investigated the effects of corporate front-group

stealth campaigns. Very similarly to astroturfing activities, these groups disseminate persuasive messages while masking their true identity and interests. After they were confronted with the disguised corporate messages, the opinions of those initially favoring restrictive policies on different issues were significantly eroded. Given the theoretical and empirical evidence, we assume that pro-Russian online comments will influence the opinions of those who read them.

H1 Exposing individuals to pro-Russian astroturfing comments will change their opinions in the direction of the comments.

The effects of astroturfing comments on opinion certainty

Whereas an attitude or opinion represents a person's evaluation of an object, situation, or person, attitude or opinion certainty refers to the extent to which one is confident in it (Gross, Holtz, & Miller, 1995). Certainty is an important dimension of attitudes and opinions, because it influences their stability, durability, and behavioral impact. There are several theoretical reasons why astroturfing comments can be expected to influence opinion certainty. Firstly, research has shown that opinion certainty can be altered by messages contradicting an existing opinion, because these decrease the structural consistency of the underlying beliefs or knowledge. Hence, information with contradictory evaluative implications should decrease opinion certainty (Smith, Fabrigar, MacDougall, & Wiesenthal, 2008). Secondly, opinion certainty is influenced by the subjective ease with which opinion-relevant information comes into an individual's mind. If information supporting the opinion is easily cognitively retrieved (e.g., because the individual has recently been exposed to it), it is deemed more valid and thus fosters opinion certainty (Tormala, Petty, & Briñol, 2002). Conversely, easily retrieved counter-attitudinal information—as provided by astroturfing comments—should decrease certainty. Finally, people hold opinions with greater certainty when they perceive social consensus for them (e.g., Visser & Mirabile, 2004). As other

studies have shown, online user comments can serve as indicators of such a consensus (Zerback & Fawzi, 2017).

Although creating uncertainty among people in democratic societies is considered a central goal of political astroturfing (Zhang et al., 2013), only the previously mentioned study by Cho and colleagues (2011) and another by Kang and colleagues (2016), who replicated the former's examination of uncertainty, have investigated such effects. Both show that exposure to astroturfing websites on global warming increases uncertainty regarding the causes of climate change and the role played by humans in this context. Based on the theoretical work and empirical studies described, we assume that counter-attitudinal astroturfing comments will decrease individual opinion certainty.

H2 Exposing individuals to pro-Russian astroturfing comments will decrease opinion certainty.

Inoculation as a countermeasure to the effects of astroturfing comments

Given the supposed effects of astroturfing comments, the question arises as to what can be done to neutralize them. One effective way to inhibit or even prevent the impact of persuasive attacks is to inoculate people against them (see Compton & Pfau, 2005). Inoculation theory explains this process by reference to a biological analogy (McGuire, 1964): resistance to future persuasive messages can be increased by administering a weakened version of the "virus" to the individual—in this case, the impending persuasive message. An inoculation procedure consists of two core elements: threat and refutational preemption (see Compton, 2012 for an overview). Threat means that the individual receives a warning about a pending persuasive attack that will challenge its attitudes. Following this warning, the person is provided with information to strengthen the existing attitude. This second element is termed *refutational preemption*, and exists in two common variants: *refutational-same preemptions* raise and refute exactly the same arguments as used in the subsequent attack message, whereas *refutational-different preemptions* include arguments that are not part of the

subsequent attack. Empirical studies have shown that both preemption types can increase resistance to attack messages (Banas & Rains, 2010; McGuire, 1964).

Despite the promising potential of the inoculation approach, to our knowledge no study to date has investigated its effectiveness in the context of astroturfing campaigns, although leading scholars in the field have emphasized its benefits and suitability to counter contemporary forms of disinformation (van der Linden, Maibach, Cook, Leiserowitz, & Lewandowsky, 2017). While some researchers have tested the effectiveness of inoculation strategies in the context of mis- or disinformation, their studies do not deal with astroturfing campaigns or state-induced propaganda in general, but rather with conspiracy theories (Banas & Miller, 2013), media reports on climate change (Cook, Lewandowsky, & Ecker, 2017), and front-group stealth campaigns (Pfau et al., 2007). Nevertheless, all these studies confirm the effectiveness of inoculation in hampering the effects of persuasive messages on personal opinions.

Whereas the works described above focused on opinion change, Tormala and Petty (2002) offer an additional perspective that also allows to derive theoretical assumptions regarding opinion certainty. They argue that the mere experience of resisting a persuasive attack can increase certainty, but only when the attack is perceived to be strong. Although the authors clearly point out the differences between the inoculation approach and their theoretical conception, they state: "As long as resistance does occur, the stronger the attack is perceived to be, the stronger the predicted effects [on certainty] will be" (p. 1300). Because an inoculation message empowers resistance to a persuasive attack, we expect increased levels of opinion certainty in those who receive an inoculation treatment. This assumption has also been confirmed by empirical studies showing that attitude certainty increased after participants were inoculated against persuasive messages (Compton & Pfau, 2004; Pfau et al., 2004). Therefore, we assume the following:

H3 Administering an inoculation treatment prior to astroturfing comments will inhibit the assumed persuasive effects on opinion change (*H3a*) and opinion certainty (*H3b*).

Durability of inoculation effects

One of the most challenging questions in the context of inoculation is how long it provides protection from persuasive messages. McGuire (1964) assumes that some time must pass between the inoculation and the attack in order to strengthen resistance. However, due to a declining motivation over time to defend one's opinion, wear-out effects may occur, decreasing resistance in the long run (Insko, 1967). The co-occurrence of both processes led researchers to assume that the effectiveness of inoculation follows an inversely U-shaped curve, which brings up the question of the ideal time interval between inoculation and attack (Compton & Pfau, 2005). Empirical studies have used varying time intervals, ranging from attack messages immediately following the inoculation treatment to intervals of several months. In their meta-analysis, Banas and Rains (2010) found some support for a declining immunizing effect when they compared short (immediate attack message), moderate (attack message after 13 days), and long (attack message after 14 days or later) intervals. However, the decline was not significant. In his literature review, Compton (2012) found some indication of a drop in resistance after a two-week period. Hence, we propose the following research question:

RQ1 Will inoculation effects on opinion change (*H3a*) and opinion certainty (*H3b*) still exist after a two-week delay between inoculation and the astroturfing comments?

Method

Our study is based on a three-wave online experiment employing a 3 (issue) x 5 (inoculation) x 2 (delay between inoculation and attack message) between subject design. Participants were recruited via a commercial online access panel (Consumer Fieldwork) in September 2018 and were randomly assigned to one of the experimental conditions. 2,353 subjects took part in all three waves of the experiment.[1] They were 48.8 years (*SD* = 15.2) old on average; 44.4 %

possessed the highest German high-school degree, and 49.9 % were female. We opted to conduct our study in the German context, primarily for reasons of political relevance. Germany is one of the EU's most powerful member states and, according to expert opinion (van Herpen, 2015), one of the prime targets of Moscow's recent propaganda efforts. In addition, we assumed that the psychological mechanisms underlying the hypothesized effects should be broadly valid across different cultural contexts (for potential risks of generalizing to other contexts, see Discussion).

Stimulus and procedure

Because online astroturfing comments often occur in the context of journalistic content (Kovic et al., 2018), our experimental stimulus consisted of a fictitious Facebook news teaser ostensibly from the largest German television newscast Tagesschau. To increase the generalizability of the results, we produced three identically designed teasers, each consisting of the Tagesschau logo, a picture, the story's headline, and a short lead text (screenshots are included in the Online Appendix). The teasers differed, however, with regard to the Russia-related issue they dealt with. Two of the issues (the murder attempt on Sergei Skripal and the manipulation of the 2016 US election) related to direct Russian involvement. The third issue (the use of toxic gas in Syria) involved the Syrian government—a close ally of Russia. Each of the three teasers blamed either the Russian (issues one and two) or the Syrian governments (issue three) for the calamity.

Furthermore, each teaser was accompanied by two user comments representing typical astroturfing attack messages. In constructing the astroturfing messages, we followed the analysis of the EU vs. Disinformation initiative, which identified the most prevalent argumentative figures used by Russian propagandists (EU vs. Disinformation, 2019). Specifically, the astroturfing comments all expressed doubt regarding Russian/Syrian involvement in the event, offering arguments supporting this position and alternative explanations. To make sure that the strength of the arguments did not differ between the

different issues, all comments were pre-tested by 17 to 20 subjects who were not part of the final study. All arguments were perceived to be moderately strong, with no significant differences between the issues (see Table 1, Online Appendix).

[FIGURE 1]

The three waves took place between September and October 2018 (Figure 1). In wave one, we measured participants' prior opinions and opinion certainty for all three issues and collected socio-demographic information. To avoid raising suspicion regarding the true goal of these questions, the first wave took place two weeks before the actual stimulus presentation. In addition, all issue-specific questions were embedded in larger item sets also encompassing other issues. Two weeks later, in wave two, participants received a second questionnaire including the inoculation treatments. In line with our theoretical outline, three different inoculation messages were administered. The *threat only* condition (IC1) included only a warning about commenters paid by the Russian government, who attempt to sway citizens' opinions regarding the respective issue. In the *refutational-different* condition (IC2), subjects received the same warning, but were additionally informed about the general persuasive strategies employed, namely, that the commenters would try to offer alternative explanations for the event in order to exonerate Russia/Syria. Subjects were also told that these alternative explanations contradicted independent official investigations of the events. Similarly, in the *refutational-same* condition (IC3), subjects were warned about the persuasive attack and informed about the strategy; however, this time by telling them the exact arguments that the commentators would use (see the Online Appendix for the inoculation messages).

In order to determine the persuasive effects of the astroturfing comments on opinions and opinion certainty (*H1* and *H2*), the inoculation factor also included two additional control conditions, in which subjects did not receive an inoculation treatment. In control condition 1 (CC1) participants were only exposed to the news teaser, in CC2 they saw the teaser

including the comments. Consequently, differences between the two control groups indicate the astroturfing comments' effects.

To assess the durability of the inoculation (*RQ1*), all subjects in wave two received the inoculation treatment; however, half of them saw the teaser including the comments immediately after the inoculation, the other half two weeks later (wave three).

Measures

Because all astroturfing comments were intended to raise doubt about Russian/Syrian involvement in the events presented, we asked our participants specifically for their opinion on the Russian/Syrian government's responsibility for the event, and how certain they were of this opinion. Subjects' *opinions* were measured using a five-point Likert scale indicating agreement with the statement that Russia/Syria was responsible for the event described in the news teaser (1 "Do not agree" to 5 "Fully agree"). The measure for *opinion certainty* was adopted from Tormala and Petty (2002), asking how certain the subjects were of the opinion indicated (1 "Not certain at all" to 5 "Extremely certain"). Subtracting participants' post-stimulus from their pre-stimulus answers, resulted in two scores, reflecting changes in opinion and opinion certainty (opinion change: $M_{Syria} = 0.24$; $SD_{Syria} = 1.04$; $M_{Skripal} = 0.33$; $SD_{Skripal} = 1.07$; $M_{US\ election} = 0.30$; $SD_{US\ election} = 1.00$; change in opinion certainty: $M_{Syria} = 0.27$; $SD_{Syria} = 1.27$; $M_{Skripal} = 0.31$; $SD_{Skripal} = 1.31$; $M_{US\ election} = 0.07$; $SD_{US\ election} = 1.21$). Positive values of the opinion-change measure indicate that respondents held Russia/Syria less responsible for the events after seeing the stimulus. Positive values of the opinion-certainty-change measure indicate higher uncertainty as compared to their initial certainty assessment.

Results

Manipulation checks

Manipulation checks yielded satisfying results. Most subjects in the inoculation conditions correctly recalled having received an inoculation message (88.5%). Likewise, most subjects

in the non-inoculation conditions correctly remembered that they had not seen such a message (87.9 %), $\chi^2(2, N = 2221) = 1281.99$, $p = .000$. Similarly, most of the participants who were exposed to astroturfing comments correctly remembered having seen comments beneath the news teaser (77.1%), as did those in the non-comment condition, where 73.8% stated that they had not seen any comments, $\chi^2(2, N = 2233) = 540.66$, $p = .000$.

Effects of online astroturfing comments on opinions and opinion certainty

To test whether the online astroturfing comments affected participants' opinions, we first focus on the two control conditions and compare participants who only saw the news teaser (CC1) to those additionally exposed to the astroturfing comments (CC2). Figure 2 depicts opinion changes in both groups (see Table 2, Online Appendix for means and statistical tests). In order to test for group differences, we followed Hayes (2005) and dummy-coded the inoculation factor. K-1 dummy variables entered a linear regression model as independents. The respective comparison group served as the reference category. Besides testing for significant mean differences, the unstandardized regression coefficient *b* indicates the direction and magnitude of the mean difference between the two groups.

Firstly, it is interesting to see that, over the course of the two weeks between the pre- and post-stimulus measurements, subjects in all issue conditions became more supportive of the Russian/Syrian position. However, while this effect was only marginal in the news-teaser-only condition (CC1) ($M = 0.12$, $SD = 0.96$), it was clearly pronounced for those who had been exposed both to the news teaser and to the online astroturfing comments ($M = 0.42$, $SD = 1.08$). Put differently, those who found pro-Russian/pro-Syrian astroturfing comments beneath the news teaser ascribed significantly less responsibility to Russia/Syria for the event, $b = 0.30$, $p < .001$. From a cross-issue perspective, *H1* can thus be confirmed. However, a closer inspection of the issue-specific patterns shows that the astroturfing comments' effect can mainly be traced back to the Skripal case, $b = 0.54$, $p = .000$, and somewhat to the Syria issue, $b = 0.21$, $p = .09$. Hence, *H1* finds support only in this case.

[FIGURES 2 and 3]

We further assumed that astroturfing comments would increase uncertainty in those who initially thought that Russia/Syria was responsible for the negative events (*H2*). Therefore, unlike in the previous analysis, we confine our examination to subjects who had initially seen the two states as culprits (indicated by values of pre-stimulus opinions of 4 or 5; $N = 995$). Figure 3 shows that, astroturfing comments affected opinion certainty in the expected direction across all issue conditions (see Table 3, Online Appendix for means and statistical tests). Again, when comparing the two control groups CC1 ($M = 0.34$, $SD = 1.19$) and CC2 ($M = 0.64$, $SD = 1.11$), participants who saw counter-attitudinal astroturfing comments became significantly more uncertain of their initial view that Russia/Syria were to blame, $b = 0.30$, $p = .009$, as compared with those who did not see the comments. Again, an issue-specific examination shows that the effect was only significant in the Skripal scenario, $b = 0.58$, $p = .005$. Therefore, *H2* can only be confirmed in this case.

Effects of inoculation treatments

In a next step, we examine whether the three inoculation strategies were able to prevent the effects of the astroturfing comments. To do so, we compare the three groups which saw the astroturfing comments after being inoculated (IC1, IC2, and IC3) to the group which received them without prior inoculation (CC2). An effective inoculation treatment should have prevented opinion change, ideally reducing it to the level of those who had only seen the news teaser without any astroturfing comments (CC1). A visual inspection of Figure 2 supports this notion, at least for the *refutational-same* inoculation treatment (IC3), $b = -0.20$, $p = .007$: participants who were educated in advance about Russia's persuasive goals and exact arguments were less influenced by the astroturfing comments ($M = 0.22$, $SD = 1.05$) as compared with non-inoculated subjects ($M = 0.42$, $SD = 1.08$). In contrast, the remaining two inoculation strategies (*threat only*: $b = -0.04$, $p = .561$; *refutational-different*: $b = -0.06$, $p = .391$) did not prevent opinion change. A further issue-specific examination of the data shows

that the overall effect of the *refutational-same* preemption was largely rooted in the Skripal and Syria cases. Multiple group comparisons indicate that the *refutational-same* strategy reduced opinion change in both issue conditions to a sufficient level, leading to a significant difference from non-inoculated participants receiving comments (CC2) (b_{Syria} = -0.24, p = .06; $b_{Skripal}$ = -0.36, p = .01) and a non-significant difference from those who had only seen the news teaser (CC1) (b_{Syria} = -0.03, p = .84; $b_{Skripal}$ = 0.18, p = .15). Hence, *H3a* finds support in these two cases (see Table 2, Online Appendix for means and statistical tests).

Following the previous logic, we finally examined the efficiency of inoculation in relation to opinion-certainty changes (*H3b*). Again, the visual patterns in Figure 3 seem to support the effectiveness of the *refutational-same* treatment, which hampered the increase in uncertainty (M = 0.43, SD = 1.19) as compared to non-inoculated subjects in CC1 (M = 0.64, SD = 1.11), although not to a highly significant extent, b = -0.21, p = .07. As Table 3 (Online Appendix) shows, none of the three inoculation strategies was able to prevent changes in opinion certainty within the single-issue conditions significantly.

Duration of inoculation effects

In a final step, we examined how long the observed immunization effect persisted (*RQ1*). The two lines in Figure 4 represent the different delay conditions implemented in our experiment (immediate and delayed astroturfing attack). It is important to recall that delay represents a between factor, so for each delay condition, we collected data across all inoculation groups.

[FIGURE 4]

As can be seen, the two lines mostly parallel each other, with only minor and non-significant differences (see Table 4 in Online Appendix for means and statistical tests). However, there is one noteworthy exception, which manifests itself in a nearly significant interaction effect between inoculation and delay, $F(4, 2054) = 2.23$, p = .06. The refutational-same treatment, which was the most potent in reducing opinion changes, was only effective when administered immediately prior to the astroturfing comments ($M_{short\ delay}$ = 0.09, $SD_{short\ delay}$ =

1.02), whereas its effect largely diminished after two weeks ($M_{long\ delay}$ = 0.36, $SD_{long\ delay}$ = 1.06), $t(385) = -2.64$, $p = .01$. When we look at the issue-specific short- and long-term effects, we find exactly the same pattern, but, again, only in the Skripal case, indicating a significant decrease over time in the immunizing effect of the refutational-same treatment ($M_{short\ delay}$ = -0.13, $SD_{short\ delay}$ = 1.10; $M_{long\ delay}$ = 0.60, $SD_{long\ delay}$ = 1.12), $t(135) = -3.18$, $p = .002$. Corresponding mean differences in the Syria condition, $t(118) = -0.233$, $p = .816$, and US election condition, $t(137) = -0.909$, $p = .365$, could not be observed.

With regard to changes in opinion certainty, we found no significant three-way interaction between issue, inoculation strategy, and delay, $F(8, 965) = 0.55$, $p = .820$. Short- and long-term inoculation effects on opinion certainty did not differ significantly across the three issue conditions.

Discussion

In this paper, we examined the persuasive effects of astroturfing comments posted beneath news items on Facebook in the context of Russia-related issues. The results show that astroturfing comments can indeed change audiences' political opinions and increase uncertainty. However, these effects did not occur equally across the issues. While we could clearly observe effects in the Skripal case and to some extent in the Syria scenario, we could not find them in the context of the manipulations of the 2016 US presidential election.

Against this background, the question arises as to what caused the issue-specific differences. We see several possible explanations. The first reason could be that participants' initial opinions were already less certain in the Skripal and the Syria conditions and therefore easier to influence by the astroturfing attacks. However, our data does not support this interpretation. On the contrary, a comparison of pre-stimulus opinion-certainty scores shows that Syria and Skripal were the cases with the highest initial certainty levels (M_{Syria} = 3.38, SD_{Syria} = 1.32; $M_{Skripal}$ = 3.31, $SD_{Skripal}$ = 1.31; $M_{US\ election}$ = 3.22, $SD_{US\ election}$ = 1.23), $F(2, 4314) = 17.984$, $p = .000$, $\eta^2 = .008$. A second possibility might be that differences in issue

involvement account for the effect patterns. Highly involved individuals are more likely to scrutinize the quality of the arguments included in a persuasive message (Petty & Cacioppo, 1984), hence the impact of the astroturfing comments could have been stronger for more involving issues. To test this explanation, we analyzed pre-stimulus involvement scores, which were measured by two items ("I think the issue is important", "I am interested in the issue") on a five point Likert scale reaching from 1 "Do not agree at all" to 5 "Totally agree". The resulting scale showed good reliability ($\alpha_{Syria} = .86$, $\alpha_{Skripal} = .88$, $\alpha_{US\ election} = .83$). A comparison of pre-stimulus involvement indeed shows significant differences between the three issues. Involvement was highest in the case of Syria ($M = 3.90$, $SD = 1.08$), followed by the US election ($M = 3.47$, $SD = 1.18$) and Skripal ($M = 3.18$, $SD = 1.18$), $F(2, 4619) = 522.282$, $p = .000$, $\eta^2 = .183$. This implies two things: first, participants perceived all issues to be at least moderately important and relevant; second, the involvement pattern does not really correspond to the issue specific effect differences observed. For example, Skripal, as the issue with the strongest astroturfing effects, was also the least involving one for our participants. Another reason for the differential effects might be that respondents' opinions about the US presidential election and Syria represented more abstract scenarios and were therefore more difficult to process, especially when someone offers alternative explanations for them. The Skripal case, on the other hand, as a more narrowly defined and concrete event, makes it easier to understand and accept possible explanations. Unfortunately, our data did not enable us to test this assumption.

In addition to the examination of the effects of astroturfing comments, this study also advances research on inoculation, being the first to transfer the approach to an online astroturfing context. As former studies have shown, inoculation can help to confer on individuals cognitive resistance to "a range of falsehoods in diverse domains such as climate change, public health, and emerging technologies" (van der Linden et al., 2017, p. 1141). Contrary to these expectations, only one strategy was effective in mitigating the persuasive

impact of astroturfing comments: when subjects were educated in advance about the exact arguments deployed by the Russians (*refutational-same*), changes in opinions and opinion certainty were prevented. However, even the immunizing effect of the *refutational-same* treatment was only short-lived and vanished almost completely after a two-week delay. This finding is in line with other inoculation studies in the context of political issues (Pfau & Burgoon, 1988).

The potentially negative effects of immunizing citizens against astroturfing comments

With regard to transferring inoculation research to the realm of astroturfing comments, perhaps the most difficult problem relates to the fact that such comments can typically not be distinguished from genuine citizens' voices (the defining element of astroturfing). This poses a dilemma because, while inoculation messages might mitigate the harmful effects of astroturfing messages (positive consequence), they might also undermine the credibility of citizen commenting in public online spaces, and of online deliberation in general (negative consequence). This potential "side-effect" of inoculation campaigns (Compton, 2012, p. 15) could only be prevented if astroturfing comments were unambiguously identifiable and distinguishable from authentic citizen comments—which will almost never be the case. Those who initiate counter campaigns will thus have to make difficult decisions as to whether, and how, citizens can and should be inoculated against political astroturfing campaigns. Rather abstract *threat-only* treatments, for instance, can be disseminated with relatively limited costs and efforts. Yet these have the disadvantage that they undermine the credibility of online citizen debate around entire political issues. Moreover, they are, according to our findings, relatively inefficient. Highly specific *refutational-same* treatments, by contrast, can be very effective in mitigating the persuasive effects of astroturfing comments, as the findings of this study indicate. They also have the advantage that they discredit only those user comments that actually convey very narrowly defined pieces of misleading and inaccurate information. The downside of *refutational-same* inoculation

treatments is, however, that they require extensive resources to tailor and administer highly issue- and argument-specific counter messages.

Practical implications: how to inoculate audiences against astroturfing comments

Our results have at least two practical implications for how governments, social networking sites, news organizations, journalists and other actors can inoculate their audiences against astroturfing comments. First, given the limited effectiveness of threat-only and refutational-different preemptions, relatively abstract inoculation messages grounded in these two strategies appear to bear little promise with regard to reducing the impact of astroturfing campaigns. Consequently, it is highly advisable to design and disseminate inoculation messages grounded in the refutational-same strategy – that is, messages that highlight the very arguments that are deployed later in the astroturfing attack. In order to create such highly issue-specific inoculation messages, publicly funded think tanks or government-sponsored counterpropaganda units like the European Union's "East Stratcom Task Force" (EU vs. Disinformation, 2019) are indispensable. These organizations are required to continuously collect and analyze information about ongoing astroturfing campaigns. Their regular reports should not only highlight those high-profile political events, which are allegedly being targeted by astroturfing actors, but also feature the key argumentative strategies deployed. This type of analysis then needs to be channeled to media and political actors, who can use it to create issue-specific inoculation messages. Secondly, the short-term nature of the effects (even of refutational-same treatments) detected in our study implies that inoculation messages are most efficient if they are presented to an audience immediately before this audience receives astroturfing comments. In practical terms, this means that the most promising strategy for administering this type of inoculation messages appears to be banners or warnings placed in the immediate vicinity of commenting fields. Media and political actors need to inoculate their audiences "just in time". At a more abstract level, with regard to media literacy campaigns in general, our findings suggest that such campaigns are

most effective when (a) designed as continuous rather than one-time efforts and (b) deploying issue-specific ("refutational-same") rather than abstract ("threat-only") messages.

Limitations and promising paths for future research

Our study also has limitations. Although we increased external validity by including three different issues in our design and by investigating the short- and long-term effects of astroturfing and inoculation messages, we still relied on results gathered in an experimental setting. Participants were purposely exposed to stimuli that they otherwise might not have encountered, for example because they did not use social media or did not read the comments beneath news articles. In this sense, the effects of comments that we found probably overestimate the effect on society as a whole. On the other hand, participants in our experiment were only exposed once to the astroturfing comments and to the inoculation messages. In a real-world environment, people probably encounter comments repeatedly, which enhances the astroturfing comments' persuasive power. The same is true of inoculation messages: simply because a one-time inoculation proves to be inefficient or loses its effect after a while, this does not mean that inoculation is an ineffective strategy. It seems plausible that multiple treatments would sustain the immunization or might even increase it by aggregating the effects of the single treatments. The question of how repeated exposure influences the persuasive effects of astroturfing comments, and those of inoculation messages, represents a promising avenue for future research.

Moreover, we conducted our experiment within one (the German) context only, based on the assumption that the basic psychological effects investigated in this study would be broadly valid across cultural contexts. Future research is needed to bolster this assumption, and our claims to generalizability, by replicating similar experimental designs in other sociopolitical settings. Context-dependent characteristics of participants that might moderate the effects reported in this study may include: participants' prior knowledge about Russia's foreign propaganda efforts, their levels of education, their general attitude towards Russia,

their political ideologies, their migration backgrounds, as well as their personal ties with Russians. Unfortunately, except for age, gender, and education, based on our data, we were not able to test for potentially moderating variables (see Table 5 in the Online Appendix).

In addition, in this study we tested our hypotheses across three issues, which were all related to Russia's involvement in political events that had occurred outside Germany. As our findings show, the persuasive impact of astroturfing comments differed greatly between the three issues. Against this backdrop, a key task for further research appears to be to incorporate a broader range of issues in future research designs, and to specify the reasons for issue-specific differences. In this context, we think the role of perceived issue relevance and involvement deserves more research attention. Particularly from a theoretical perspective, issue involvement is an interesting factor, because it is assumed to moderate simultaneously the effects of persuasive (Petty & Cacioppo, 1984) and inoculation messages (Pfau et al., 1997). Moreover, in the context of inoculation, involvement has not only been shown to serve as an independent or moderating variable, but can also be the result of inoculation messages (Compton & Pfau, 2004). Developing an integrative theoretical framework describing the complex role of involvement in both contexts would be a promising task for future research.

Finally, the astroturfing comments used, were designed as grounded in only one – even though arguably the most prevalent – argumentative technique deployed in Russia's recent disinformation campaigns: that of denying the principal's (Russia's) responsibility for a negative event, and of offering alternative explanations. Going beyond this study, future research could explore how audiences can most efficiently be inoculated against other common propaganda techniques, such as pointing to a general "Russophobia" or discrimination against Russians, ironizing the accusations, or relativizing the breaches of norms ("everybody does this") (EU vs. Disinformation, 2019). By following up on these and related paths, future research can theorize and investigate in significantly more depth the

mechanisms that facilitate the disconcerting persuasive potential of disinformation and discover promising strategies for minimizing its harmful effects on democratic life.

References

Banas, J. A., & Miller, G. (2013). Inducing resistance to conspiracy theory propaganda: Testing inoculation and metainoculation strategies. *Human Communication Research*, *39*(2), 184–207. https://doi.org/10.1111/hcre.12000

Banas, J. A., & Rains, S. A. (2010). A meta-analysis of research on inoculation theory. *Communication Monographs*, *77*(3), 281–311. https://doi.org/10.1080/03637751003758193

Bennett, W. L., & Livingston, S. (2018). The disinformation order: Disruptive communication and the decline of democratic institutions. *European Journal of Communication*, *33*(2), 122–139. https://doi.org/10.1177/0267323118760317

Bessi, A., & Ferrara, E. (2016). Social Bots Distort the 2016 US Presidential Election Online Discussion. *First Monday*, *21*(11). Retrieved from https://ssrn.com/abstract=2982233

Bugorkova, O. (2015). Ukraine conflict: Inside Russia's 'Kremlin troll army'. *BBC News*. Retrieved from https://www.bbc.com/news/world-europe-31962644

Cho, C. H., Martens, M. L., Kim, H. [Hakkyun], & Rodrigue, M. (2011). Astroturfing global warming: It isn't always greener on the other side of the fence. *Journal of Business Ethics*, *104*(4), 571–587. https://doi.org/10.1007/s10551-011-0950-6

Compton, J. A. (2012). Inoculation theory. In J. Dillard & L. Shen (Eds.), *The SAGE Handbook of Persuasion: Developments in Theory and Practice* (pp. 1–20). Thousand Oaks, CA: Sage Publications. https://doi.org/10.4135/9781452218410.n14

Compton, J. A., & Pfau, M. (2004). Use of inoculation to foster resistance to credit card marketing targeting college students. *Journal of Applied Communication Research*, *32*(4), 343–364. https://doi.org/10.1080/0090988042000276014

Compton, J. A., & Pfau, M. (2005). Inoculation theory of resistance to influence at maturity: Recent progress in theory development and application and suggestions for future research. *Annals of the International Communication Association*, *29*(1), 97–146. https://doi.org/10.1080/23808985.2005.11679045

Cook, J., Lewandowsky, S., & Ecker, U. K. H. (2017). Neutralizing misinformation through inoculation: Exposing misleading argumentation techniques reduces their influence. *PloS One*, *12*(5), 1-21. https://doi.org/10.1371/journal.pone.0175799

Daschmann, G. (2000). Vox pop & vox polls: The impact of poll results and voter statements in the media on the perception of a climate of opinion. *International Journal of Public Opinion Research*, *12*(2), 160–181. https://doi.org/10.1093/ijpor/12.2.160

Domke, D., Shah, D. V., & Wackman, D. B. (1998). Media priming effects: Accessibility, association, and activation. *International Journal of Public Opinion Research*, *10*(1), 51–74. https://doi.org/10.1093/ijpor/10.1.51

EU vs. Disinformation (2019). Conspiracy mania marks one-year anniversary of the Skripal poisoning. Retrieved from https://euvsdisinfo.eu/conspiracy-mania-marks-one-year-anniversary-of-the-skripal-poisoning/

European Commission. (2018). *A multi-dimensional approach to disinformation: Report of the independent high level group on fake news and online disinformation*. Luxembourg: Publications Office of the European Union.

Ferrara, E. (2017). Disinformation and social bot operations in the run up to the 2017 French presidential election. *First Monday*, *22*(8), 1-30. https://doi.org/10.2139/ssrn.2995809

Gross, S. R., Holtz, R., & Miller, N. (1995). Attitude certainty. In R. E. Petty & J. A. Krosnick (Eds.), *Attitude strenght. Antecedents and consequences* (4th ed., pp. 215–245). Mahwah, NJ: Lawrence Erlbaum Associates.

Hayes, A. F. (2005). *Statistical methods for communication science*. Mahwah, N.J: Lawrence Erlbaum Associates.

Hovland, C. I., Janis, I., & Kelley, H. H. (1953). *Communication and persuasion: Psychological studies of opinion change*. New Haven, CO, London: Yale University Press.

Insko, C. A. (1967). *Theories of attitude change*. New York: Appleton-Century-Crofts.

Kang, J., Kim, H. [Hyungsin], Chu, H., Cho, C. H., & Kim, H. [Hakkyun] (2016). In distrust of merits: The negative effects of astroturfs on people's prosocial behaviors. *International Journal of Advertising*, *35*(1), 135–148. https://doi.org/10.1080/02650487.2015.1094858

Keller, F. B., Schoch, D., Stier, S., & Yang, J. (2019). Political astroturfing on Twitter: How to coordinate a disinformation campaign. *Political Communication*, *63*(2), 1–25. https://doi.org/10.1080/10584609.2019.1661888

King, G., Pan, J., & Roberts, M. E. (2017). How the Chinese government fabricates social media posts for strategic distraction, not engaged argument. *American Political Science Review*, *111*(03), 484–501. https://doi.org/10.1017/S0003055417000144

Kovic, M., Rauchfleisch, A., Sele, M., & Caspar, C. (2018). Digital astroturfing in politics: Definition, typology, and countermeasures. *Studies in Communication Sciences*, *18*(1), 69–85.

Lefevere, J., Swert, K. de, & Walgrave, S. (2012). Effects of popular exemplars in television news. *Communication Research*, *39*(1), 103–119. https://doi.org/10.1177/0093650210387124

Lysenko, V., & Brooks, C. (2018). Russian information troops, disinformation, and democracy. *First Monday*, *23*(5).

McGuire, W. J. (1964). Inducing resistance to persuasion: Some contemporary approaches. In L. Berkowitz (Ed.), *Advances in experimental social psychology* (1st ed., pp. 191–229). New York: Academic Press.

McNutt, J., & Boland, K. (2007). Astroturf, technology and the future of community mobilization: Implications for nonprofit theory. *The Journal of Sociology & Social Welfare*, *34*(3), 165–178.

Nimmo, B. (2015). Anatomy of an info-war: How Russia's propaganda machine works, and how to counter it. Retrieved from https://www.stopfake.org/en/anatomy-of-an-info-war-how-russia-s-propaganda-machine-works-and-how-to-counter-it/

Petty, R. E., & Cacioppo, J. T. (1984). The effects of involvement on responses to argument quantity and quality: Central and peripheral routes to persuasion. *Journal of Personality and Social Psychology*, *46*(1), 69–81. https://doi.org/10.1037/0022-3514.46.1.69

Pfau, M., & Burgoon, M. (1988). Inoculation in political campaign communication. *Human Communication Research*, *15*(1), 91–111. https://doi.org/10.1111/j.1468-2958.1988.tb00172.x

Pfau, M., Compton, J. A., Parker, K. A., Wittenberg, E. M., An, C., Ferguson, M., . . . Malyshev, Y. (2004). The traditional explanation for resistance versus attitude accessibility. *Human Communication Research*, *30*(3), 329–360. https://doi.org/10.1111/j.1468-2958.2004.tb00735.x

Pfau, M., Haigh, M. M., Sims, J., & Wigley, S. (2007). The influence of corporate front-group stealth campaigns. *Communication Research*, *34*(1), 73–99. https://doi.org/10.1177/0093650206296083

Pfau, M., Tusing, J. K., Koerner, A. F., Lee, W., Godbold, L. C., Penaloza, L. J., . . . Hong, Y.-H. (1997). Enriching the inoculation construct: The role of critical components in the process of resistance. *Human Communication Research*, *24*(2), 187–215. https://doi.org/10.1111/j.1468-2958.1997.tb00413.x

Ruck, D. J., Rice, N. M., Borycz, J., & Bentley, R. A. (2019). Internet Research Agency Twitter activity predicted 2016 U.S. election polls. *First Monday*, *24*(7). https://doi.org/10.5210/fm.v24i7.10107

Sikorski, C. von (2018). The effects of reader comments on the perception of personalized scandals: Exploring the roles of comment valence and commenters' social status. *International Journal of Communication*, *10*, 4480–4501.

Smith, S. M., Fabrigar, L. R., MacDougall, B. L., & Wiesenthal, N. L. (2008). The role of amount, cognitive elaboration, and structural consistency of attitude-relevant knowledge in the formation of attitude certainty. *European Journal of Social Psychology*, *38*(2), 280–295. https://doi.org/10.1002/ejsp.447

Taylor, S. E., & Thompson, S. C. (1982). Stalking the elusive "vividness" effect. *Psychological Review*, *89*(2), 155–181. https://doi.org/10.1037//0033-295X.89.2.155

Tormala, Z. L., & Petty, R. E. (2002). What doesn't kill me makes me stronger: The effects of resisting persuasion on attitude certainty. *Journal of Personality and Social Psychology*, *83*(6), 1298–1313. https://doi.org/10.1037/0022-3514.83.6.1298

Tormala, Z. L., Petty, R. E., & Briñol, P. (2002). Ease of retrieval effects in persuasion: A self-validation analysis. *Personality and Social Psychology Bulletin*, *28*(12), 1700–1712. https://doi.org/10.1177/014616702237651

Van der Linden, S., Maibach, E., Cook, J., Leiserowitz, A., & Lewandowsky, S. (2017). Inoculating against misinformation. *Science*, *358*(6367), 1141–1142. https://doi.org/10.1126/science.aar4533

Visser, P. S., & Mirabile, R. R. (2004). Attitudes in the social context: The impact of social network composition on individual-level attitude strength. *Journal of Personality and Social Psychology*, *87*(6), 779–795. https://doi.org/10.1037/0022-3514.87.6.779

Weedon, J., Nuland, W., & Stamos, A. (2017). *Information operations on Facebook*. Retrieved from https://fbnewsroomus.files.wordpress.com/2017/04/facebook-and-information-operations-v1.pdf

Woolley, S. C., & Guilbeault, D. R. (2017). *Computational propaganda in the United States of America: Manufactoring consensus online*. Working Paper No. 2017.5. Oxford: University of Oxford.

Zelenkauskaite, A., & Balduccini, M. (2017). "Information warfare" and online news commenting: Analyzing forces of social influence through location-based commenting user typology. *Social Media + Society*, *3*(3), 1-13. https://doi.org/10.1177/2056305117718468

Zerback, T., & Fawzi, N. (2017). Can online exemplars trigger a spiral of silence? Examining the effects of exemplar opinions on perceptions of public opinion and speaking out. *New Media & Society*, *19*(7), 1034–1051. https://doi.org/10.1177/1461444815625942

Zerback, T., & Peter, C. (2018). Exemplar effects on public opinion perception and attitudes: The moderating role of exemplar involvement. *Human Communication Research*, *14*(2), 125. https://doi.org/10.1093/hcr/hqx007

Zhang, J., Carpenter, D., & Ko, M. (2013). *Online Astroturfing. A theoretical perspective: Proceedings of the Nineteenth Americas Conference on Information Systems, Chicago, Illinois, August 15-17, 2013.*

Zillmann, D. (1999). Exemplification theory: Judging the whole by some of its parts. *Media Psychology*, *1*(1), 69–94. https://doi.org/10.1207/s1532785xmep0101_5

---

[1] A statistical power analysis showed that, in order to find small interaction effects between all three experimental factors, a minimum sample size of $N = 2{,}283$ is necessary.

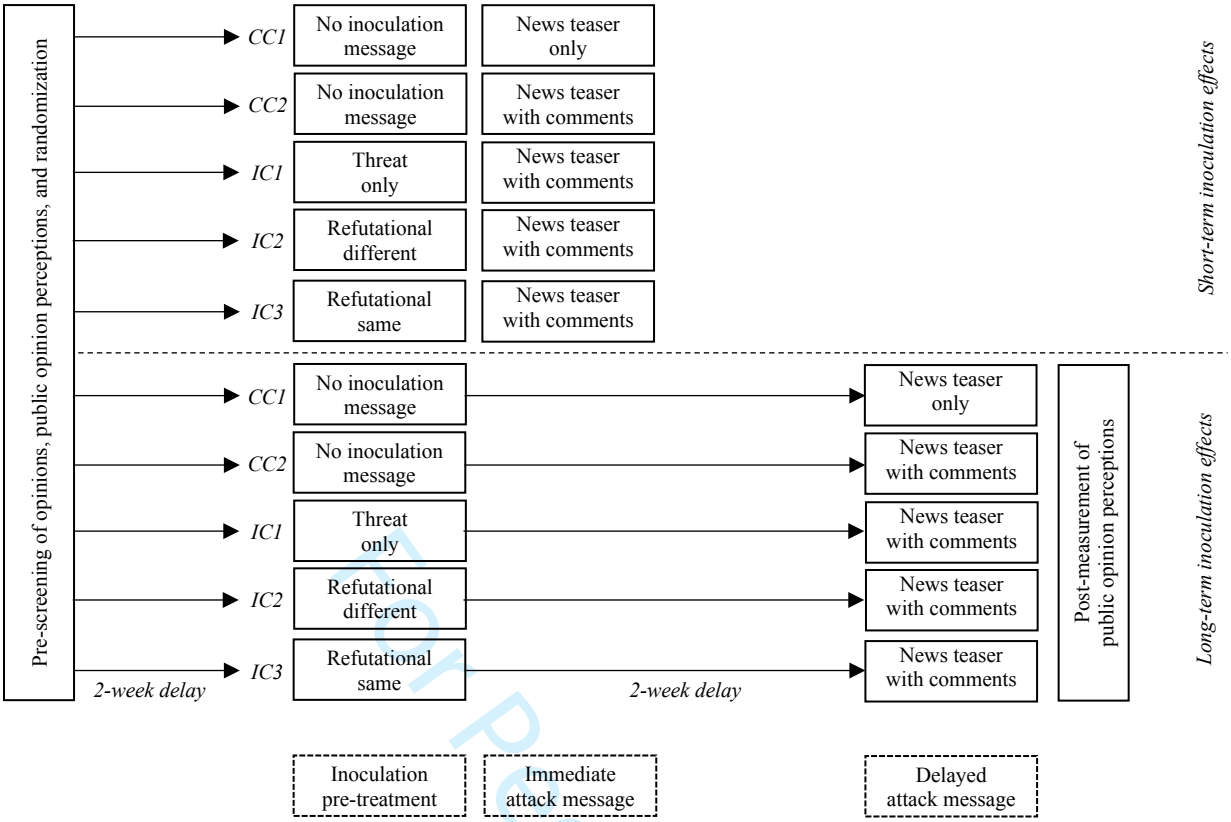


Figure 1

Experimental design

The design was replicated for all three issue conditions.

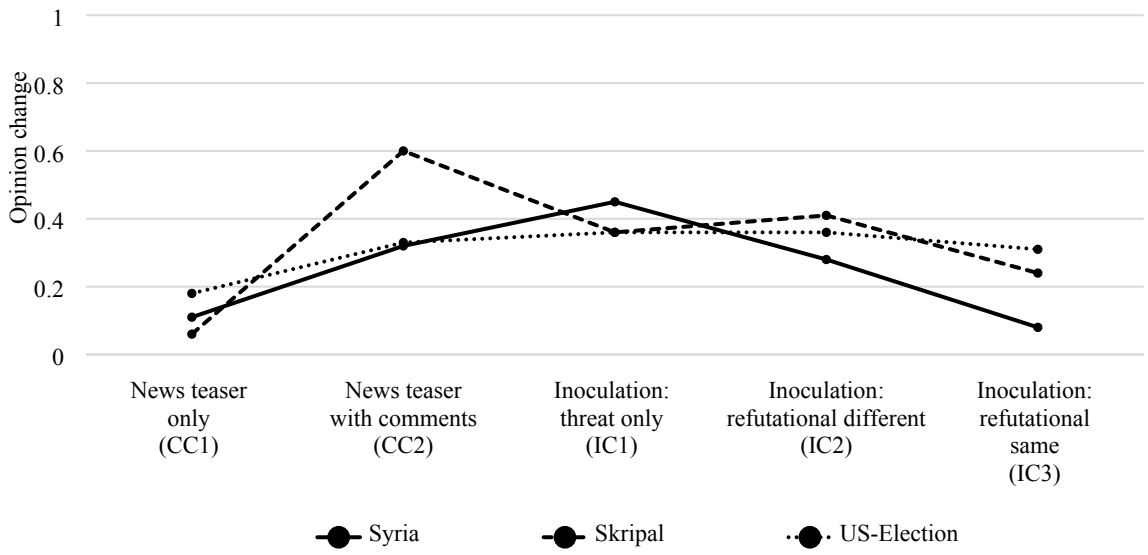


Figure 2

Effects of astroturfing comments and inoculation treatments on opinion change

$N = 2{,}064$

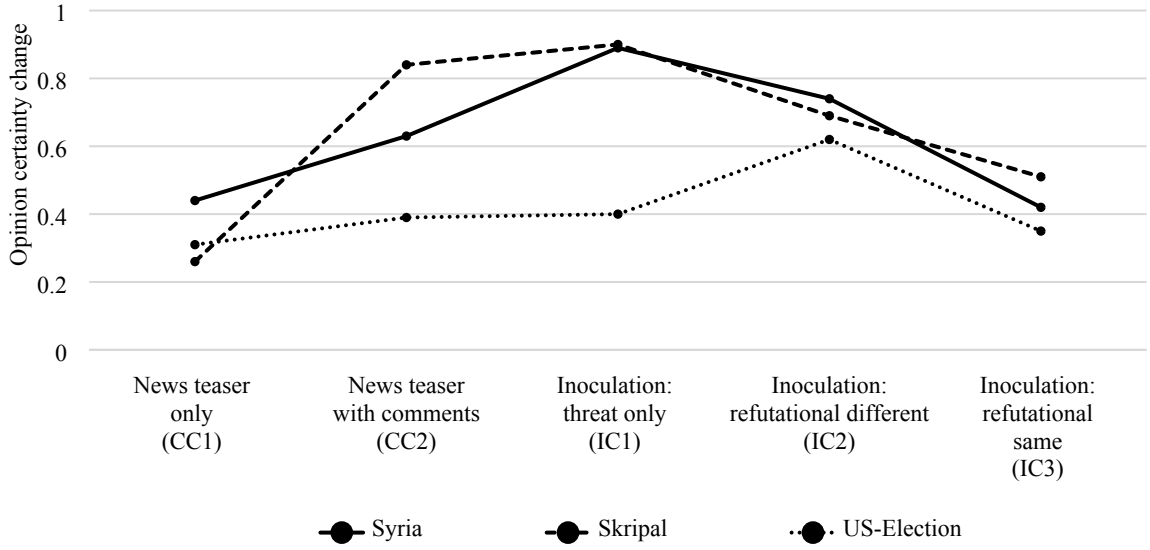


Figure 3

Effects of astroturfing comments and inoculation treatments on opinion-certainty change

N = 995 participants initially indicating that Russia/Syria was responsible for the event depicted (values of pre-stimulus opinion 4 or 5).

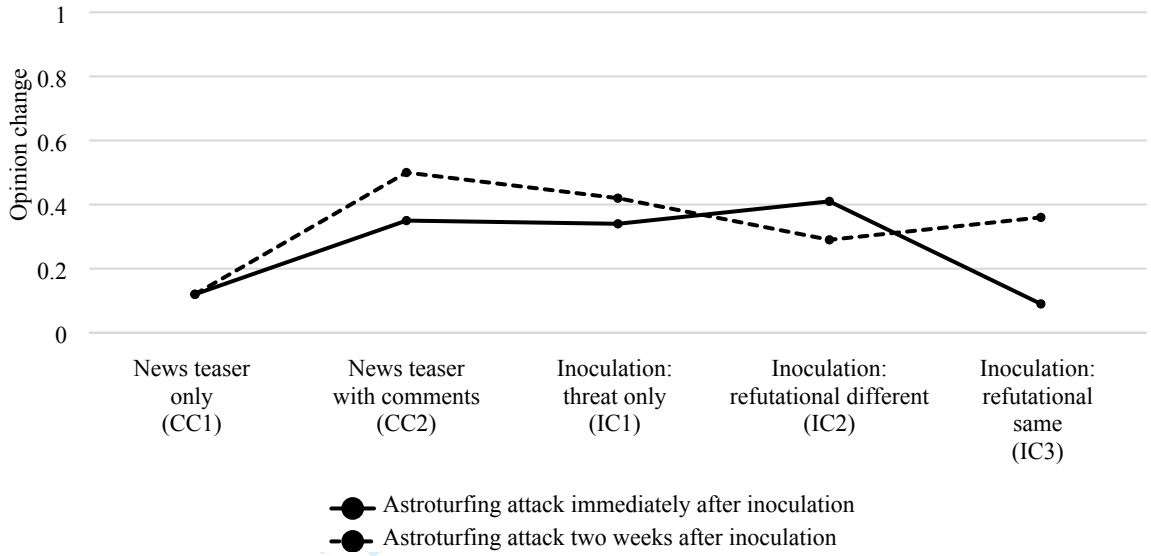


Figure 4

Short- and long-term inoculation effects on opinion change

N = 2,064

Materials Supplementary to Article

The disconcerting potential of online disinformation: Persuasive effects of astroturfing comments and three strategies for inoculation against them

Contents

Table 1 Perceived argument strength in online astroturfing comments (pretest results)

	Argument strength	
	M (SD)	*α*
Syria (*N* = 19)		
Comment 1: *"Like Assad's the only one with poison gas. What about the thousands of IS henchmen? If someone is known for massacring civilians, then it's probably them."*	3.71 (0.85)	0.91
Comment 2: *"Nothing's proved! Wouldn't be the first time somebody invented weapons of mass destruction to wage a fucking war."*	3.59 (1.21)	0.97
Overall strength	3.65 (0.94)	0.94
Skripal (*N* = 20)		
Comment 1: *"So the guy was a proven double agent and had connections to the mafia. There were a lot of other people who wanted to kill him."*	2.73 (1.18)	0.96
Comment 2: *"If the Russians wanted Skripal dead, they simply would have done it without leaving traces. But no, they used a poison that directly points to them. Right!"*	3.01 (1.41)	0.96
Overall strength	2.87 (1.21)	0.96
US election (*N* = 17)		
Comment 1: *"Russian wire-pullers? Yeah sure! Cold-blooded economic interests are behind the election manipulations: Facebook, Cambridge Analytica. Do I need to say any more?"*	2.65 (1.02)	0.93
Comment 2: *"Nothing's proved! It wouldn't be the first time someone manipulated an election to gain power in the country."*	3.32 (1.02)	0.95
Overall strength	2.99 (0.81)	0.91

* *N* = 58 participants took part in the pretest and indicated argument strength on a bipolar scale from 1 to 5 using the following items: not convincing – convincing, weak – strong, implausible – plausible, incorrect – correct. All items were used to construct a scale indicating the perceived strength of each argument.

Table 2 Group mean differences in opinion change

	Teaser only (CC1)		Teaser with astroturfing comments (CC2)		Inoculation: Threat only (IC1)		Inoculation: Refutat.-different (IC2)		Inoculation: Refutat.-same (IC3)	
	M (SD)		*M (SD)*		*M (SD)*		*M (SD)*		*M (SD)*	
Syria (n = 657)	0.11[c]	(1.02)	0.33	(1.12)	0.40[ae]	(1.01)	0.28	(0.96)	0.08[c]	(1.04)
Skripal (n = 685)	0.06[bcd]	(0.83)	0.60[ae]	(1.15)	0.36[a]	(1.21)	0.41[a]	(0.94)	0.24[b]	(1.14)
US election (n = 722)	0.18	(1.03)	0.33	(0.98)	0.36	(1.08)	0.36	(0.99)	0.31	(0.95)
All issues (N = 2,064)	0.12[bcd]	(0.96)	0.42[ae]	(1.08)	0.37[ae]	(1.10)	0.35[ae]	(0.96)	0.22[bc]	(1.05)

Group comparisons are based on linear multiple regression analysis using the inoculation factor as a dummy variable. Superscripts indicate significant mean differences ($p < .05$) between the groups (a = Teaser only, b = Teaser with astroturfing comments, c = Inoculation: Threat only, d = Inoculation: Refutational-different, e = Inoculation: Refutational-same).

Table 3 Group mean differences in opinion certainty change

	Teaser only (CC1)		Teaser with astroturfing comments (CC2)		Inoculation: Threat only (IC1)		Inoculation: Refutat.-different (IC2)		Inoculation: Refutat.-same (IC3)	
	M (SD)		*M (SD)*		*M (SD)*		*M (SD)*		*M (SD)*	
Syria (n = 331)	0.44[c]	(1.26)	0.63	(1.12)	0.89[ae]	(1.18)	0.74	(0.98)	0.42[c]	(1.32)
Skripal (n = 349)	0.26[bcd]	(1.26)	0.84[a]	(1.24)	0.90[a]	(1.34)	0.69[a]	(1.12)	0.51	(1.28)
US election (n = 315)	0.31	(1.05)	0.39	(0.84)	0.40	(1.07)	0.62	(1.22)	0.35	(0.96)
All issues (N = 995)	0.34[bcd]	(1.19)	0.64[a]	(1.11)	0.75[ae]	(1.23)	0.68[ae]	(1.11)	0.43[cd]	(1.19)

*Participants stating that Russia / Syria was responsible for the event depicted (pre-stimulus opinions 4 or 5). Positive values indicate higher opinion uncertainty. Group comparisons are based on linear multiple-regression analysis using the inoculation factor as a dummy variable. Superscripts indicate significant mean differences ($p < .05$). (a = Teaser only, b = Teaser with astroturfing comments, c = Inoculation: Threat only, d = Inoculation: Refutational-different, e = Inoculation: Refutational-same).

Table 4 Differences in opinion change after short and long delays

	Opinion change									
	Teaser only (CC1)		Teaser with astroturfing comments (CC2)		Inoculation: Threat only (IC1)		Inoculation: Refutat.-different (IC2)		Inoculation: Refutat.-same (IC3)	
	M (SD)		*M (SD)*		*M (SD)*		*M (SD)*		*M (SD)*	
Short delay (n =1,107)	0.12	(0.97)	0.35	(1.14)	0.34	(1.12)	0.41	(0.96)	0.09[a]	(1.02)
Long delay (n = 957)	0.12	(0.95)	0.50	(0.99)	0.42	(1.08)	0.29	(0.96)	0.36[a]	(1.06)

Group comparisons represent simple main effects of delay. Superscripts indicate significant mean differences between short- and long-delay conditions within a single inoculation group ($p < .05$).

Table 5 Moderation of astroturfing effects by age, education, and gender

	Attitude change				Attitude certainty change			
	Syria	Skripal	US election	Overall	Syria	Skripal	US election	Overall
Age[1]	0.001 n.s.	-0.015 n.s.	-0.008 n.s.	-0.007 n.s.	-0.006 n.s.	-0.025 n.s.	0.003 n.s.	-0.009 n.s.
Education[2]	0.003 n.s.	0.012 n.s.	0.007 n.s.	0.001 n.s.	0.000 n.s.	0.001 n.s.	0.001 n.s.	0.000 n.s.
Gender[2]	0.002 n.s.	0.017 n.s.	0.005 n.s.	0.000 n.s.	0.006 n.s.	0.003 n.s.	0.002 n.s.	0.003 n.s.

[1]Unstandardized coefficient of interaction term (astroturfing*age) within linear regression model.

[2]Partial Eta² of interaction terms (astroturfing*gender, astroturfing*education) within ANOVA model.

$^{***}p < .001$, $^{**}p < .01$, $^{*}p < .05$

Figure I Inoculation messages (Skripal issue)

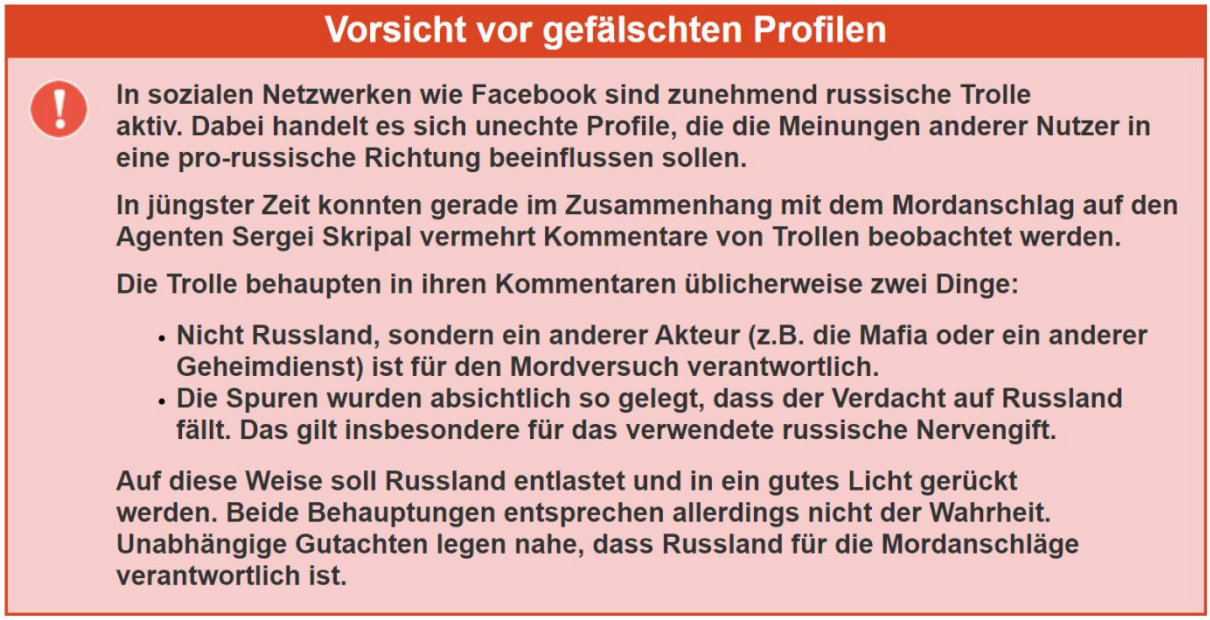
Thread only

Translation: [Title] Beware of fake profiles [Text] Russian trolls are increasingly active on social networks like Facebook. They use fake profiles to sway the opinions of other users in a pro-Russian direction. Recently, troll comments have been frequently observed, particularly in the context of the assassination attempt on agent Sergei Skripal.

Refutational-different

Translation: [Title] Beware of fake profiles [Text] Russian trolls are increasingly active on social networks like Facebook. They use fake profiles to sway the opinions of other users in a pro-Russian direction. Recently, troll comments have been frequently observed, particularly in the context of the assassination attempt on agent Sergei Skripal. In their comments, the trolls often spread alternative explanations for the events to . In this way, Russia is to be to exculpated and put in a good light. However, their explanations do not correspond to the truth.

Refutational-same



Translation: [Title] Beware of fake profiles [Text] Russian trolls are increasingly active on social networks like Facebook. They use fake profiles to sway the opinions of other users in a pro-Russian direction. Recently, troll comments have been frequently observed, particularly in the context of the assassination attempt on agent Sergei Skripal. The trolls usually state two things in their comments:

- Not Russia, but another actor (e.g. the mafia or another secret service) is responsible for the attempted murder.
- The clues were deliberately placed so that Russia is suspected. This is particularly applies to the Russian nerve toxin that was used.

In this way, Russia is to be to exculpated and put in a good light. However, their explanations do not correspond to the truth. Independent reports suggest that Russia is responsible for the murder attempt.

Figure II Issue-specific news teasers (including astroturfing comments)

Skripal issue

Translation: [Title] Who poisoned ex-spy Skripal? Multiple countries claim it was Russia. [Caption] Russian secret service suspected [Comment 1] So the guy was a proven double agent and had connections to the mafia. There were a lot of other people who wanted to kill him. [Comment 2] If the Russians wanted Skripal dead, they simply would have done it without leaving traces. But no, they used a poison that directly points to them. Right!

Syria issue

Translation: [Title] In the past year, more than 150 people were killed during gas attacks in Syria. Multiple countries claim it was Assad. [Caption] Syrian government suspected [Comment 1] Like Assad's the only one with poison gas. What about the thousands of IS henchmen? If someone is known for massacring civilians, then it's probably them. [Comment 2] Nothing's proved! Wouldn't be the first time somebody invented weapons of mass destruction to wage a fucking war.

US election issue

Translation: [Title] Did Russia manipulate the US presidential election? The country is in the focus of ongoing investigations [Caption] Russia suspected [Comment 1] Russian wire-pullers? Yeah sure! Cold-blooded economic interests are behind the election manipulations: Facebook, Cambridge Analytica. Do I need to say any more? [Comment 2] Nothing's proved! It wouldn't be the first time someone manipulated an election to gain power in the country.